Revisiting State Legislative Campaign Finance and Elections: Illinois and Texas

Selina Wu

April 2020

```
cf <- read.csv("campaign_finance.csv")
```

Difference in average Democratic vote share in Illinois vs. in Texas

An aggregate function is used to calculate the average Democratic vote shares of state, then the difference between the averages of IL and TX are calculated.

```
state_DVS <- aggregate(cf$dem_vote_pct, list(cf$state), mean, na.rm=T)
colnames(state_DVS) <- c("State", "Mean_DVS")

avgTX <- state_DVS[state_DVS$State=="TX",2]
avgIL <- state_DVS[state_DVS$State=="IL",2]

avgIL-avgTX

## [1] 7.510541
```

95% confidence interval for difference-in-means (without using lm() or t.test())

The standard error must be calculated in order to determine the confidence interval. The degrees of freedom from this data is extremely high at 31776, so the t statistic used can be substituted for the z-score when alpha/2 = 0.025.

```
meanDVS <- mean(cf$dem_vote_pct, na.rm=T)
diff <- state_DVS$Mean_DVS - meanDVS
avg_diff <- mean(diff, na.rm=T)

nrow(cf)

## [1] 31777

se <- sqrt( var(state_DVS[state_DVS$State,2])/nrow(state_DVS))
diff_CI <- c(avg_diff - (1.96*se), avg_diff + (1.96*se))
diff_CI

## [1] -2.543378  1.917965
```

In practice, we can use the t-distribution even if the Democratic vote share is not normally distributed. As n, the sample size, gets bigger, the Central Limit Theorem kicks in. Thus, a normal distribution can be approximated.

Interpretation of the data and the confidence interval

The difference between the Democratic vote share in IL and TX is statistically significiant. The difference of 7.511 lies outside of the 95% confidence interval. This interval means that in repeated sampling, 95% of the intervals constructed this way will contain the true population mean. In other words, we are 95% confident that a difference in means value from the samples will lie within the given bounds. Since the difference is greater than the upper bound of the confidence interval, it is considered statistically significant.

We have enough evidence to reject the null hypothesis that IL and TX voters are equally Democratic. Since the difference lies above the upper bound of the confidence interval, we can have say that IL voters are more Democratic than those of TX.

Regression predicting Democratic vote percentage based on Democratic monetary contribution percentage

```
lm(cf$dem_vote_pct~cf$dem_money_pct)

##
## Call:
## lm(formula = cf$dem_vote_pct ~ cf$dem_money_pct)
##
## Coefficients:
##      (Intercept)  cf$dem_money_pct
##           16.102             0.689
```

For every 1% increase in overall money contribution, the democratic vote share increases by 0.689%. We can say the value for the slope is a drawn from a sample, and want to test how likely this slope is due to random chance. By constructing a confidence interval telling us if the 0.689 slope value is statistically significant, we can determine whether the relationship between DVS and money contribution is statistically significant.

90% confidence interval for the slope coefficient

A similar method as the one used in Question 4 to calculate the confidence interval is used. A null hypothesis would be that the slope value = 0, so we can use this as the mean slope value. If repeatedly taking elections as samples, 90% of the confidence intervals constructed will return slope values within this confidence interval.

```
se2 <- sqrt(var(cf[, 8])/nrow(cf))
# t(alpha/2) = t(alpha=0.05) = 1.64 as DoF approaches infinity

slope_CI <- c(-1.64*se2, 1.64*se2)
slope_CI

## [1] -0.2836785  0.2836785
```

A slope coefficient falling higher than the upper bound of the confidence interval indicates a higher association between contributions and vote share.

The slope value being statistically significant does not prove campaign contributions buy votes. It may indicate there is a relationship between the two variables, but we cannot claim causality. Campaign contributions can predict who wins an election, but this can be due to better candidates raising more money from being a high quality candidate. In order to to prove a causal relationship, we'd have to randomize campaign contributions.